DOCUMENT RESUME

ED 095 201 TM 003 873

AUTHOR Rakow, Ernest A.
TITLE Evaluation of Educational Program Differences Via Achievement Test Item Difficulties.
SPONS AGENCY Carnegie Corp. of New York, N.Y.
PUB DATE [Apr 74]
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; *Comparative Analysis; Complexity Level; Correlation; Individual Differences; *Item Analysis; Norm Referenced Tests; *Program Evaluation
IDENTIFIERS *Item Difficulty

ABSTRACT

An approach for clearer observation of differences when evaluating educational programs is presented. The standardized tests utilized in evaluating programs are designed to measure topics commonly taught and to maximize individual differences. This masks between-program differences and the unique aspects of different programs. Analysis of item difficulties for an achievement test relative to the program mean difficulty assists in identifying program strengths and weaknesses. The correlations of programs (as variables) using item difficulties as observations indicate the degree of communality between programs. These techniques should assist in evaluation of innovative educational programs. (Author)

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

7.09
P.9

ED 095201

EVALUATION OF EDUCATIONAL PROGRAM DIFFERENCES VIA ACHIEVEMENT TEST ITEM DIFFICULTIES*

Ernest A. Rakow, Boston College

BEST COPY AVAILABLE

Purpose

The purpose of this paper is to present a unique approach for more clearly observing differences when evaluating educational programs. Generally, standardized achievement tests are the central instrument utilized in evaluation of educational programs such as Title I evaluations, Project Follow-Through, and Equality of Educational Opportunity. Standardized achievement tests are designed to maximize individual differences and measure them reliably. Such tests measure topics taught in many educational programs and avoid those topics which occur in few educational programs. These tests are refined via statistical analysis of the items using the item difficulty and discrimination. These statistical procedures are applied to large samples of students from many educational programs to maximize the reliability of the measurement of individual differences. This also tends to increase measurement of common topics and avoidance of unique topics. Consequently, while standardized tests provide excellent measures of individual differences on common topics, they may not be appropriate for program evaluation. Program evaluation should focus on differences between programs, i.e. the ways in which particular programs are unique, as well as indicating adequacy on the common topics.

---

*Paper presented at the Annual Meeting of American Educational Research Association, Chicago, Illinois, April, 1974. Work on this paper was supported by a grant from Carnegie Corporation to Dr. George Madaus and Dr. Peter Airasian.

TM 003 873

BEST COPY AVAILABLE

Measurement of the unique aspects of an educational program would require testing procedures which could identify homogenous performance within a particular program but heterogeneous performance between programs. This paper presents an approach for further inspection of achievement test data when evaluating programs.

Method

The techniques presented here are further analyses of item difficulties for items of a norm-referenced achievement test. One could view each item as another observation taken on each of the programs. The mean of the item difficulties for each program could be examined to observe overall differences in level of achievement. This would yield the same conclusions as evaluating programs via the means of the test scores. Programs could also be considered as variables enabling the calculation of correlation coefficients of these observations (items) on programs (variables). If there are no program differences, other than level of achievement, these correlations should all be approximately equal and their magnitude should be close to that of the reliability of the test. However, if some correlations are considerably lower than the test reliability this is a clear indication of program uniquenesses. If there are no unique program effects then the difficulty of an item relative to the mean difficulty for a program should be approximately equal in all programs. If there are program differences, then an item which measures a unique aspects of the program should be of greater relative difficulty in one program than in another. It should be noted that this approach concentrates on performance relative to the mean of that program, not on the overall level of performance as reflected in program means.

If correlations significantly below the reliability coefficient are observed, one could proceed by further evaluating the item difficulties. One could calculate the intraclass correlation for each item between the programs.

The items with higher intraclass correlations indicate items measuring greater program differences. One could also examine the item difficulties, searching for items which may cause the lower correlation between programs.

Data

The data for this paper are the item difficulties for a 69 item mathematics achievement test administered to mathematics students in their final year of secondary school in twelve countries. There are twelve item difficulty estimates for each item, one for each country included in the sample. These item difficulties were published in a bulletin by the International Association for the Evaluation on Educational Achievement (IEA).

Results

The results of applying these procedures to the data from the twelve countries are interesting. First of all, there are significant differences in the mean level of achievement (as was reported by IEA). But that is not the point of this paper.

Treating the item difficulties as the observations on twelve countries a correlation matrix was calculated. These correlations are given in Table 1. At the bottom of this table the country means and reliabilities are also given. The median reliability for this test in these countries is 0.88.

Excluding the main diagonal, the median correlation in this matrix is 0.70, which is significantly below the lowest reliability (.79). Only six correlations are larger than the lower of the two reliability estimates for the corresponding pair of countries. Three of these high correlations are for the countries of England, Scotland and Australia (the only countries of the British Commonwealth included here). These high correlations indicate similar patterns of item difficulties relative to the country means.

One explanation for this could be similarities in the educational system and especially in the emphesis of the mathematics curriculum. Note that this is a very different interpretation than suggested by the overall means. (England is high while Australia is low.) It is also granted that a competing explanation for these high correlations could be the cultural and social similarities of these countries. The other three high correlations are for the countries of Holland, Sweden and Finland. Once again, the two competing explanations are (1) cultural and social similarities or (2) similarities in the educational system and in the empheses in the teaching of mathematics.

Fifty of the sixty-six correlation coefficients in this matrix are significantly below the lower of the two reliability estimates for that pair of countries. (Significance is defined as having a Z score for the correlation more than 1.65 standard errors below the Z score for lower of the reliability estimates. Hays, 1963.) The country with the lowest correlations is Israel. The correlations of Israel with other countries range from a low of 0.25 (the lowest in the matrix) to a high of 0.70. Other countries in which every correlation is significantly below the reliability estimates are the United States, Belgium and France. The lower correlations are the result of differences in the pattern of item difficulties relative to the mean. This would seem to indicate that the organization and emphasis on topics within the mathematics curriculum has some unique aspects for these four countries. This seems reasonable when one is aware that these tests included item testing both higher and lower mental process scores and the topics of new mathematics, elementary and intermediate algebra, Euclidian and analytic geometry, calculus, analysis and set theory. Perhaps it should be noted that the test mean for Israel is very high while that for the United States is low, so that this is more than merely indicating level of performance.

These significantly lower correlations led to further examination of the item difficulties. The next step was to calculate the intraclass correlation

for each item. This statistic provides an indication of the between country heterogeniety relative to the within group homogeniety for each item. The intraclass correlation can be interpreted as a proportion of explained variance. If there are between group differences these intraclass correlations would be greater than zero for each item. Also, if the relative performance of these items was the same, then the intraclass correlation should be approximately equal for all items. Table 2 shows this is not the case. These intraclass correlations range from a low of .020 to a high of .271.

Table 2 presents only a subset of this further analysis of the item difficulties. Only thirty of the items are presented here. These items are the ten with the highest intraclass correlations, ten with the lowest, and the middle ten. This table also provides the percentage of correct responses to these items for six of the countries and for all twelve countries combined. The last line in the table is the percentage correct on the total test of sixty-nine items. The second last line is the percentage correct on the subset of the ten items with the highest intraclass correlations. Comparison of these percentages for 69 items and for 10 items reveals the percentages for the United States and Australia are even lower for ten items than for sixty-nine items while for Israel the reverse is true, i.e., the percentage is even higher for ten items than for sixty-nine items. This is simply an indication that those ten items are more sensitive to between country differences than is the entire test. For the items with highest intraclass correlation the typical range for the percentage of correct responses for these items is about 60. For the middle ten items on the intraclass correlation the typical range is 33. For the lowest ten the typical range in percentages is 28.

This led to further analyses of the item difficulties. The percentages right on individual items for each country were compared with the percentage for all countries combined. In general, one would expect the three countries

with lower means to have items with lower percentages correct and would expect higher percentages correct in the three countries with higher means. This tends to be true. In Table 2 plus (+) signs are used in the three low countries to indicate item difficulties above the level for all twelve countries combined. For the ten items with the highest intraclass correlation, in the United States only one item has a + and there are only two +'s for Australia. Perhaps this caused the percent right for the first ten items in each of these countries to be lower than for the entire test. These are the items which were even more difficult than expected in these countries. Negative (-) signs are used in the three high countries to indicate item difficulties which are below the level for all twelve countries. For these same ten items there is only one negative for Israel. This contributed to a higher percentage right on these ten items than in the total test.

Further analysis of item difficulties within these sets of three countries could be pursued. For example, the item with the intraclass correlation of .216 (rank of 3) has a percentage right of 3 in the United States and 42 in Finland. For this same item England had 16 percent right while Israel had 62 percent. These two pairs of countries have similar means, so this item appears to indicate differences in mathematics ability which is not shown in the country means. Other items also could reveal such differences, such as the one with a rank of six or a rank of eight. On item six, the percent right for the United States is 31 while it is 4 for Finland. On item eight in England the percent correct is 70 while in Israel it is only 36. These two comparisons are the reverse of that shown in item three. The effect of combining these items would be to show little difference in performance for each of these pairs of countries. However, the item difficulties clearly indicate there is a difference.

Importance

These results indicate that analysis of item difficulties can be an important

technique in evaluating educational programs. These procedures aid in identifying unique aspects of a program which may be different from another program. Such uniquenesses may be hidden by examination of test scores and differences between means. Thus, these techniques should be an important aid in program evaluation.

TABLE 1. CORRELATIONS OF COUNTRIES ON THE BASIS OF ITEM DIFFICULTIES FOR A MATHEMATICS ACHIEVEMENT TEST

	U.S.	Bel.	Fra.	Jap.	Sco.	Aus.	Fin.	Ger.	Eng.	Swe.	Hol.	Isr.
United States	1.00											
Belgium	.73*	1.00										
France	.66*	.72*	1.00									
Japan	.65*	.65*	.76*	1.00								
Scotland	.73*	.61*	.65*	.72*	1.00							
Australia	.74*	.62*	.69*	.77*	.95#	1.00						
Finland	.62*	.68*	.75*	.82	.77*	.79*	1.00					
Germany	.61*	.54*	.65*	.69*	.77*	.79	.80	1.00				
England	.61*	.54*	.62*	.67*	.89#	.88#	.72*	.81	1.00			
Sweden	.57*	.60*	.70*	.80*	.80*	.82	.88#	.81	.78*	1.00		
Holland	.50*	.50*	.66*	.73	.71	.75	.86#	.76	.66*	.81#	1.00	
Israel	.25*	.44*	.42*	.58*	.59*	.61*	.63*	.55*	.64*	.70*	.69*	1.00
Means	25.6	37.1	36.0	37.3	31.6	27.5	30.5	33.4	39.8	32.0	34.9	42.2
Reliabilities	.92	.91	.91	.92	.86	.87	.87	.85	.92	.90	.79	.82

* Correlation is significantly below the relibility coefficient.

\# Correlation is larger than the relibility coefficient.

TABLE 2. ITEM STATISTICS: INTRACLASS CORRELATIONS AND ITEM DIFFICULTIES

Rank	Intraclass Correlation	Percentage Correct: Twelve Countries	U.S.	Aust.	Finl.	Japan	Engl.	Israel
1	.271	64	25	67+	70+	48-	91	90
2	.226	29	13	11	48+	52	28	90
3	.216	18	3	7	42+	29	16-	62
4	.215	48	39	29	75+	79	44-	75
5	.207	41	11	47+	17	31-	81	71
6	.203	23	31+	10	4	8-	29	50
7	.189	20	14	10	4	59	12-	66
8	.189	54	58	25	72+	85	70	36-
9	.188	55	20	45	63+	49-	75	82
10	.178	23	19	16	9	21-	68	43
30	.085	87	84	90+	94+	92	92	90
31	.083	52	34	34	51	54	65	58
32	.081	49	43	29	45	67	48	91
33	.080	28	16	18	24	43	40	43
34	.079	72	48	62	73	82	83	88
35	.074	71	54	62	80+	90	80	64
36	.070	41	47+	32	47+	53	49	36-
37	.069	43	23	38	53+	52	57	87
38	.069	48	29	46	44	53	63	83
39	.068	47	32	34	47	54	55	53
60	.038	66	51	59	62	71	79	77
61	.038	63	57	53	56	64	76	86
62	.037	67	49	71+	55	77	82	89
63	.035	17	25+	13	11	25	15-	1-
64	.035	29	33+	23	11	30	38	25-
65	.031	31	37+	25	14	28-	46	27-
66	.021	60	53	53	74+	61	70	55-
67	.021	21	21	15	24+	33	19	20
68	.021	62	63	64	58	55-	76	60
69	.020	51	57+	44	37	49	58	70
First 10		37.5	23.4	26.2	40.6+	46.0	51.2	66.5
All 69	.199	46.9	37.0	39.8	44.1	54.1	57.7	61.2

\+ These percentages are higher than for all twelve countries combined even though the means suggest lower percentages.

\- These percentages are lower than for all twelve countries combined even though the means suggest higher percentages.